Data Science Toolbox

Owen McDevitt

January 2019

- Measurement: gaining insight into qualitative phenomena through data

- Statistical Programming Languages: using code to run statistical analysis on a computer (Scripted languages are more user friendly, while compiled languages deal directly with the CPU and thus are faster and more powerful)

- Web Scraping: downloading data from publically available places on the internet

- Handling Large Datasets: using RDDs and Spark or Hadoop to make large datasets more managebale while preventing any loss of data due to a computer failing

- Visualization: creating visualizations of the data using a graphics package to see what the data looks like (distribution, trends, correlations, outliers, etc.)

- Modeling: using statistics to test, predict, or explain phenomena or theory